

UNCLASSIFIED



Australian Government
Department of Defence
Defence Science and
Technology Organisation

A Short Guide to Experimental Design and Analysis for Engineers

Edward H. S. Lo, T. Andrew Au and Peter J. Hoek

Joint and Operations Analysis Division
Defence Science and Technology Organisation

DSTO-TN-1291

ABSTRACT

An experiment is a test under controlled conditions to investigate the validity of a hypothesis. Experimentation is the basis for creating new knowledge – determining whether some factor causes an effect. Well-designed experiments provide a systematic approach to observe relations among variables while ruling out alternative explanations. This report has examined a number of experimental designs including the simple experiment, matched-pairs, repeated-measures and the single group design. The relevant statistical techniques are also discussed to help identify key quantitative methods for data processing and analysis. While not intending to replace any textbook, this guide summarises the resources and some worked examples that have helped the authors gain a basic understanding of design, measurement and statistical analysis to support military experiments.

RELEASE LIMITATION

Approved for public release

UNCLASSIFIED

UNCLASSIFIED

Published by

*Joint and Operations Analysis Division
DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend, Victoria 3207 Australia*

*Telephone: 1300 333 362
Fax: (03) 9626 7999*

*© Commonwealth of Australia 2014
AR-015-938
May 2014*

APPROVED FOR PUBLIC RELEASE

UNCLASSIFIED

UNCLASSIFIED

A Short Guide to Experimental Design and Analysis for Engineers

Executive Summary

An experiment is a test under controlled conditions to investigate the validity of a hypothesis. Experimentation is the basis for creating new knowledge – determining whether some factor causes an effect. The basic premise of this evidence-based approach is to discard subjectivity of authorities, and to seek the facts from scientific observation of certain phenomena. Well-designed experiments provide a systematic approach to observe relations among variables while ruling out alternative explanations. In order for an experiment to establish cause-and-effect, the experiment must have internal-validity – that is, setting up the conditions that allow a treatment's effects to be isolated. While uncontrolled experiments such as those conducted in the field do not mitigate all confounding factors, they nonetheless provide external-validity to support the results from controlled experiments.

This report has examined a number of experimental designs including the simple experiment, matched-pairs, repeated-measures and the single group design. The relevant statistical techniques are also discussed to help identify key quantitative methods for data processing and analysis. While not intending to replace any textbook, this guide summarises the resources and some worked examples that have helped the authors gain a basic of design, measurement and statistical analysis to support military experiments.

UNCLASSIFIED

UNCLASSIFIED

This page is intentionally blank

UNCLASSIFIED

Authors

Edward H. S. Lo

Joint and Operations Analysis Division

Dr Edward Lo joined DSTO in 1999 after graduating from the University of Western Australia. While working, he completed a PhD in Electrical Engineering from the University of New South Wales. In his thesis he developed novel scale and rotation invariant texture features for automatic image segmentation and query by example. He is currently a research scientist in Joint and Operations Analysis Division. His research interests include studying socio-technical systems, experimentation and applying modelling and simulation to examine C2 environments.

T. Andrew Au

Joint and Operations Analysis Division

Dr Andrew Au is a senior research scientist in the Joint Headquarters Analysis Discipline within the Joint and Operations Analysis Division of DSTO. His research interests include modelling methodologies for socio-technical systems and the development of capabilities to support command and control.

Peter J. Hoek

Joint and Operations Analysis Division

Dr Peter Hoek holds a Bachelor of Information Technology, Master of Science, and a Doctor of Information Technology. Prior to working in DSTO he was employed in the engineering of automation and remote control solutions for the mining industry. He currently works as an analyst within the Joint and Operations Analysis Division of DSTO. His research interests include effective techniques for visualisation and improved understanding of information.

UNCLASSIFIED

This page is intentionally blank

UNCLASSIFIED

Contents

1. INTRODUCTION.....	1
2. EXPERIMENTAL DESIGN	1
2.1 Simple Experiment	2
2.2 Matched-Pairs Design	4
2.3 Repeated-Measures Design (Within-Subjects Design)	5
2.4 Single Group Designs.....	5
2.5 Exploratory / Uncontrolled Experiments	6
3. MEASURING DATA AND DESCRIPTIVE STATISTICS.....	6
3.1 Descriptive Statistics	7
3.2 Visualising Results.....	7
1.1.1 Radar / Spider Chart	8
1.1.2 Box and Whisker Plots	8
1.1.3 100% Stacked Bar.....	9
1.1.4 Comparing Distributions.....	9
4. HYPOTHESIS TESTING AND INFERENTIAL STATISTICS.....	11
4.1 Central Limit Theorem.....	12
4.2 Z-Statistic	13
4.3 Standard <i>t</i> -Test.....	15
4.4 Paired <i>t</i> -Test.....	17
4.5 ANOVA	19
4.6 Two-Way ANOVA.....	24
4.7 MANOVA	27
4.8 Mann-Whitney <i>U</i> -Test / Wilcoxon Signed-Rank Test.....	30
5. SUMMARY.....	32
6. REFERENCES	33

UNCLASSIFIED

DSTO-TN-1291

This page is intentionally left blank

UNCLASSIFIED

1. Introduction

Science is derived from the Latin term *scientia*, which means “knowledge”. Accordingly, *scientific method* literally means the “method that searches after knowledge.” Scientific method originated around the 16th century when people found that when data was assembled and examined without bias, some previously undiscovered meaning might be revealed. Traditionally, scientific method seeks to understanding the unknown by (Leedy and Ormrod, 2005):

1. identifying a problem to solve
2. establishing a hypothesis that if confirmed resolves the problem
3. gather data relevant to the hypothesis
4. analysing and interpreting the data to determine whether the hypothesis is supported or not.

Experimentation is a central aspect of scientific method and is the basis for creating new knowledge - determining whether some factor causes an effect.

The Logic of Warfighting Experiments (Kass, 2006) and GUIDEx (Kass et al., 2006) provide quintessential reading for planning military experiments. However, both books have deliberately omitted some aspects of experimental design and statistics. Being from a different background (engineering), the authors have decided to collaborate on this short guide to help identify key topics on experimental design, measurement and analysis.

2. Experimental Design

An experiment is a scientific procedure used to test a hypothesis, answer a question, or prove a fact. Two common types of experiments are simple experiments and controlled experiments. A simple experiment is a specific type of study to establish a cause-and-effect and often used to determine the effect of a treatment. Experiments can be extremely complex and involve a multitude of variables. In a complex setting, a controlled experiment is considered a better experiment because it is harder for other factors to influence the results, which could lead to an incorrect conclusion.

In order for an experiment to establish cause-and-effect, the experiment must have internal-validity – that is, setting up the conditions that allow a treatment’s effects to be isolated (Mitchell and Jolley, 2010). A simple experiment provides the easiest way to establish a cause-and-effect relationship, but this approach is not always possible or viable due to real world constraints. This section details a range of experimental designs and the requirements for each to establish causality.

2.1 Simple Experiment

The goal of science is to establish and advance knowledge. Experiments provide a systematic procedure for scientists to observe relations among variables while ruling out alternative explanations (Nock et al., 2008). In human-in-the-loop experiments, ideal or simple experiments are those where participants are randomly assigned to either a treatment or control (comparison) group. The idea that some variable has an effect on another is tested in a simple experiment by considering whether measurements from a group given the treatment are statistically different to those from the control group. Examples of simple experiments where participants are randomly assigned into groups include:

- clinical trials testing the effectiveness of a drug where one group is given the treatment and the control group given a placebo,
- tests to establish whether some widget improves performance of X where one group is given the new device and the control group employing the standard technique

or

- experiments to examine whether a process change is more efficient than the existing process where groups are assigned to one method or another.

While random assignment mitigates selection bias as an explanation for differences between treatments (Slavin, 2007), it doesn't produce identical groups. Results from each group will have statistical variation due to chance (i.e. different group means). Testing for statistical significance determines whether the difference between groups is due to something other than random error. As a rule of thumb, simple experiments should employ at least 30 participants in each condition to minimise error due to differences between participants (Mitchell and Jolley, 2010).

Table 1 summarises the possible outcomes of a statistical significance decision from an experiment. A correct decision is made when the decision from statistical significance testing agrees with the actual state of affairs. A type 1 error occurs when the analysis reports a treatment as having an effect when in fact it does not, that is differences due to chance are mistaken for real differences. In contrast, a type 2 error occurs when a treatment is reported as not having an effect when in reality there is. Under these circumstances, a treatment did have an effect but the study failed to detect it (Mitchell and Jolley, 2010).

Table 1: Possible outcomes of statistical significance decisions (Mitchell and Jolley, 2010)

Statistical Significance Decision	True State of Affairs	
	Treatment has an Effect	Treatment doesn't have an Effect
Significant: Reject the Null Hypothesis	Correct decision	Type 1 error
Not significant: Don't reject Null Hypothesis	Type 2 error	Correct decision

Descriptive statistics are techniques for describing a sample and include measures of central tendency (mean, median and mode), frequency distributions and graphs. Comparisons using this approach on data from two groups reveal differences between those specific groups but the **results cannot be generalised to a larger population**. Importantly, comparing the group's performances by comparing means this way to determine the treatment effect neglects to account for random variations due to chance. To **account for these variations and for conclusions from the experiment to be applied more generally require techniques from inferential statistics** such as the *t*-test and ANOVA. These and other techniques are described in Section 4 under Hypothesis Testing.

Results from each group will be spread about some mean value. These variations are termed random errors and may be due to (Mitchell and Jolley, 2010):

- random measurement error
- random differences between testing situations
- random differences between participants
- data entry errors when coding data.

These all contribute to type 2 errors. Ways to mitigate these errors to design powerful experiments include taking greater care in data entry, standardising procedures, using reliable measures, increasing numbers of trials or using a homogeneous group of participants. Another way to reduce type 2 errors is to create a larger treatment effect such as by setting higher doses or increasing difficulty levels.

Also known as false positives, type 1 errors occur when a chance difference is mistakenly considered for a real difference. There is only one way to deal with type 1 errors and that is to decide on the risk of making such an error through significance level, α . The concept of statistical significance is most easily understood by an example. Suppose a coin used in coin tossing is being assessed for bias because previous flips revealed a tendency for heads to appear. An experiment to test for bias is conducted by ten flips of the coin with the results compared against the chances for heads to appear in such circumstances for a fair coin. The chances of obtaining 8, 9 or 10 heads in 10 tosses from a fair coin is shown in Table 2. As shown scoring ten heads out of ten tosses is not impossible, albeit a very small chance.

Table 2: Probability occurring from 10 coin tosses (Mitchell and Jolley, 2010).

Event	Probability occurring from 10 coin tosses			
	Expression	%	Decimal	Calculations
8 or more heads	$P(X \geq 8) = P(X = 8) + P(X \geq 9)$	5.47%	0.0547	$\binom{10}{8} \left(\frac{1}{2}\right)^8 \left(1 - \frac{1}{2}\right)^{10-8} + P(X \geq 9)$
9 or more heads	$P(X \geq 9) = P(X = 9) + P(X = 10)$	1.07%	0.0107	$\binom{10}{9} \left(\frac{1}{2}\right)^9 \left(1 - \frac{1}{2}\right)^{10-9} + P(X = 10)$
10 heads	$P(X = 10)$	0.1%	0.001	$\binom{10}{10} \left(\frac{1}{2}\right)^{10} \left(1 - \frac{1}{2}\right)^{10-10}$

The level of significance, α determines the level of risk of making type 1 errors when making a conclusion. For the coin toss example, setting α to 0% (or 0.0) implies accepting no risk in making type 1 errors, meaning that not even ten heads from ten tosses will be convincing enough to declare a bias coin because the p -value (chance) of 0.001 is larger than α . Setting $\alpha = 0.05$ or 5% implies considering the coin biased when scoring nine or more heads from ten tosses, but not with eight or more heads. Reducing the risk of making type 1 error increases the probability of type 2 errors because genuine treatment effects (such as actually using a biased coin) are overlooked (Mitchell and Jolley, 2010).

Declaring a statistically significant result implies beyond reasonable doubt that differences observed between groups is due to the treatment and not chance. Here, reasonable doubt is taken to be some value of α that is typically set to 0.05 or 5%. For this reason, we often see a declaration that “the results were statistically significant ($p < 0.05$)” (Mitchell and Jolley, 2010).

2.2 Matched-Pairs Design

While a simple experiment provides one of the easiest ways of determining whether a factor causes an effect, practical limitations often prevent participants from being randomly assigned to each group. In some cases, randomly assigning at least 30 participants from a population to each group poses a significant challenge that prevents use of the simple experiment. A matched-pairs design is an alternative approach combining the best attributes of matching and random assignment that requires fewer participants than with simple experiments (Mitchell and Jolley, 2010).

Whereas a simple experiment attempts to minimise random errors between groups by increasing numbers of participants, a matched-pairs design seeks to do so by creating control and treatment groups with similar attributes. Similar groups are created by measuring participants with respect to a variable correlating with the dependent measure. For example, participants in a memory experiment would be given a memory test that then allows them to be ranked according to their scores. Paring the two highest scores and repeating this step for the remaining ranked participants, then randomly assigning one

member of each pair to either the control or treatment group produces the similar groups (Mitchell and Jolley, 2010).

2.3 Repeated-Measures Design (Within-Subjects Design)

The ultimate way to eliminate random error due to individual differences is to employ the same group for both treatment and control. In repeated-measures designs, each participant receives all types of treatments administered in the experiment and measured after each type of treatment. A restriction with a between-subjects design such as a matched-pairs design or simple experiment is the limit of one observation per participants. Within-subjects designs however allow getting at least two observations per participant. For example, in a study of women's ratings of men based on masculinity, Frederick and Haselton had participants performing octuple duty – providing ratings of attractiveness for eight drawings that varied in muscularity (Mitchell and Jolley, 2010). However, a potential risk with a repeated-measures design relates to order (trial) effects due to taking multiple observations from the same participant. That is, in a within-subjects design the treatment may not be the only factor being manipulated due to order effects being a confounding factor. Mitchell and Jolley (2010) identify four sources of order effects:

Practice: learning from earlier treatments improves performance in subsequent trials

Fatigue: performance decreases due to tiredness in subsequent trials

Treatment carryover: effects of earlier treatments carry to responses in latter trials

Sensitisation: participants realise what the independent and dependent variables in latter parts of the experiment and may act to support the hypothesis rather than reacting to the treatment

Randomising the order of trials is one way of balancing out order effects, while giving participants extensive practice before the experiment minimises practice effects. Treatment carryover effects can be reduced by allowing sufficient time for the effects from a previous treatment to wear off. Reducing the demands of a treatment or shortening the duration helps minimise fatigue effects while minimising sensitisation effects might involve preventing participants from noticing the changes between treatments.

2.4 Single Group Designs

Whilst some Defence human-in-the-loop experiments employ a single operator such as for evaluating cockpit design or a command support system, command and control (C2) experiments typically involve a unit (group) where it is difficult to allocate a unique team to each trial due to a small pool of participants. For this reason, the same unit receives all treatment conditions in a single group design (Kass et al., 2006). As in repeated-measures designs, order effects are a risk for establishing cause-and-effect in single group designs because latter trials may be positively biased due to learning or memory effects or negatively biased from fatigue. Kass et al. suggest using a counterbalanced design to reduce order effects (see Table 3).

Table 3: A counterbalanced experimental design.

Mon	Tue	Wed	Thu
Current	Future	Future	Current

If units are presented with identical scenarios for all conditions, the appropriate statistical analysis technique is the paired *t*-test or repeated-measures (M)ANOVA (Gueorguieva and Krystal, 2004). Otherwise, if scenarios between conditions are similar but not exactly the same, then the *t*-test or (M)ANOVA becomes the appropriate method for analysis.

2.5 Exploratory / Uncontrolled Experiments

Defence experiments such as the study of autonomous workgroups in dynamic targeting in Exercise Pitch Black 2008 (Lo et al., 2013) or the CAGE series (Lo et al., 2014) of joint fires activities have been uncontrolled in nature. An uncontrolled experiment does not mitigate confounding factors, thus preventing statements about cause-and-effect from being determined. In CAGE3A (run in 2013), issues included:

1. order effects (memory, fatigue and learning effects)
2. controlled variables were not held constant (IT systems regularly failing)
3. independent variables were not held constant under the same condition (removing players, removing monitors, introducing unplanned systems, changing seating from days 2 or 3 in the To-Be week)
4. multiple hypotheses being tested in a single experiment.

Ideally, an experiment should only test a single hypothesis. According to Neuman (2006), an experiment is rarely appropriate for research questions that require studying the impact of dozens of diverse variables simultaneously. Rarely do experiments enable assessing conditions across a wide range of complex settings or numerous social groups all at the same time. During CAGE3A, causes from one hypothesis appeared to impact another, meaning that an alternative cause not identified in the hypothesis was responsible for the effect. For this reason, uncontrolled experiments only produce anecdotal evidence supporting or refuting experimental objectives (Lo et al., 2014). However, this is not to say that uncontrolled experiments aren't important. For example, field experiments as exploratory experiments examining interventions in the real world can provide valuable insights into the sociotechnical system, identify problems to investigate or gain new insights to develop hypotheses. A campaign of controlled and uncontrolled experiments demonstrates internal and external validity for cause-and-effect statements.

3. Measuring Data and Descriptive Statistics

Studies of sociotechnical systems invariably involve gauging people's behaviours, characteristics, attitudes and opinions. Techniques to facilitate the quantification and

evaluation of possibly complex behaviours and attitudes include the checklist and the rating scale. A checklist is a list of behaviours or characteristics under investigation that allows the researcher or participant to tick off if observed, present or true or vice-versa. A rating scale is useful when a behaviour, attitude or opinion needs to be evaluated on an interval scale of measurement such as in Table 1. Developed by Rensis Likert, rating scales are sometimes referred as Likert Scales (Leedy and Ormrod, 2005). A Likert scale has the advantage of producing responses as quantitative data reflecting degrees of opinion.

Table 4: Likert scales can generate quantitative metrics from a variety of questions¹

	(Low)	Likert Scale			(High)
	1	2	3	4	5
Agreement	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
Frequency	Never	Rarely	Occasionally	Frequently	Very Frequently
Importance	Unimportant	Of Little Importance	Moderately Important	Important	Very Important
Likelihood	Almost Never True	Usually Not True	Occasionally True	Usually True	Almost Always True

3.1 Descriptive Statistics

Once collated, performance between specific groups can be compared using descriptive statistics for describing a sample including measures of central tendency (mean, median and mode), frequency distributions and graphs. Measures of variability include dispersion and deviation while a measure of relationship is correlation. However, correlation doesn't necessarily indicate causation (Leedy and Ormrod, 2005). Results from descriptive statistics cannot be generalised to a larger population. Generalising to a larger population requires inferential statistics as discussed in Section 4.

3.2 Visualising Results

An alternative approach to comparing numeric results through descriptive statistics is through visualisation. This section surveys some notable approaches including,

1. Radar / Spider Chart
2. Box and Whisker Plots
3. 100% Stacked Bar
4. comparing estimates of density functions fitted from the data.

Choice of the approach is dependent on the type of input data (naturally qualitative versus those derived from a Likert scale) or the number of dependent variables.

¹ S. A. McLeod (2008), *Likert Scale – Simply Psychology*. <http://www.simplypsychology.org/likert-scale.html>, accessed 4 April 2014

1.1.1 Radar / Spider Chart

A radar or spider chart such as shown in Figure 1 allows one set of multi-dimensional data to be compared against another (Few, 2005). In a C2 experiment, average performance across a number of variables such as communication, leadership, team workload, shared situation and others may be compared against mean scores after the application of a treatment. Another use of radar charts is in comparing performance over the days for a specific characteristic such as timeliness.

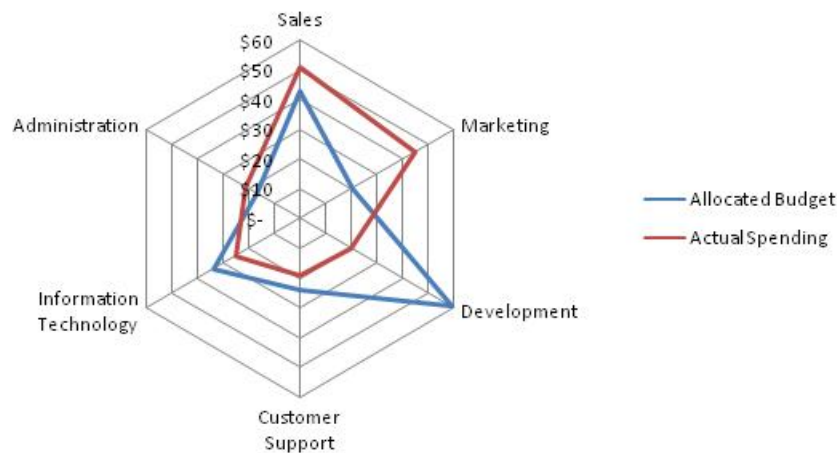


Figure 1: An example radar chart comparing sets of multi-dimensional data²

1.1.2 Box and Whisker Plots

A box and whisker plot aggregates a box plot showing the first, second and third quartile levels with whiskers representing another pair of user-defined large and small values (Mergerdichian et al., 2012). Popular values presented by whiskers include:

- mean value \pm one standard deviation,
- maximum / minimum values sampled
- 2nd and 98th percentile.

Box plots provide a quick way to assess statistical differences between samples for a specific measure. Placed side-by-side, a pair of box and whisker plots provide a graphical way to compare statistics of a measure from two or more samples (see example in Figure 2).

² Wikipedia Radar Chart http://en.wikipedia.org/wiki/Radar_chart, accessed July 2013

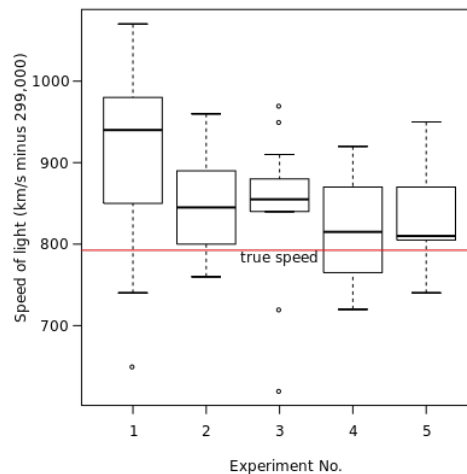


Figure 2: Examples of box and whisker plots³

1.1.3 100% Stacked Bar

A 100% stacked bar enables comparison of distributions within categories. Each row of the graph represents all responses such as from Likert scales for that category. By matching responses across rows according to common ratings (such as Strongly Disagree, Disagree, Agree, etc.) a visual comparison can be made across categories (see Figure 3).

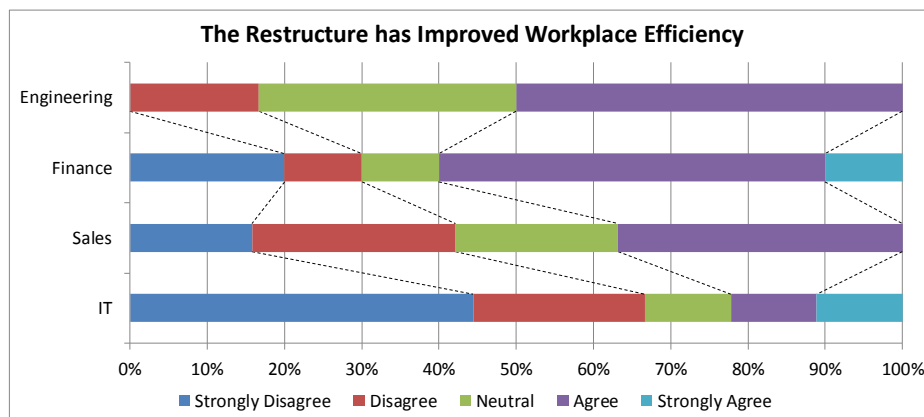


Figure 3: Example of a 100% stacked bar.

1.1.4 Comparing Distributions

If the form of the distribution is known, the parameters may be evaluated algebraically or in some cases, solved numerically from histograms generated from the set of measurements. For example, solving the parameters for a Normal distribution simply involves evaluating the mean and standard deviation from the sample, while doing so for a Rayleigh distribution requires non-linear optimisation. Standard techniques for the latter include the Gauss-Newton, the Levenberg-Marquardt and Powell's Dog Leg method (Madsen et al., 2004).

³ Wikipedia Box Plot http://en.wikipedia.org/wiki/Box_plot, accessed July 2013

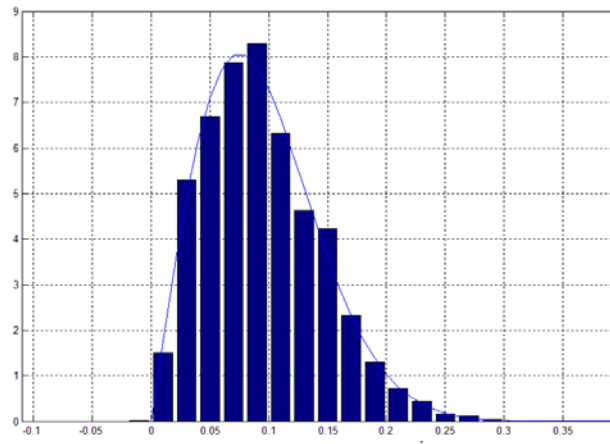


Figure 4: Fitting a Rayleigh distribution over a histogram.

Once fitted, the resulting distributions may be compared side-by-side as shown in Figure 5.

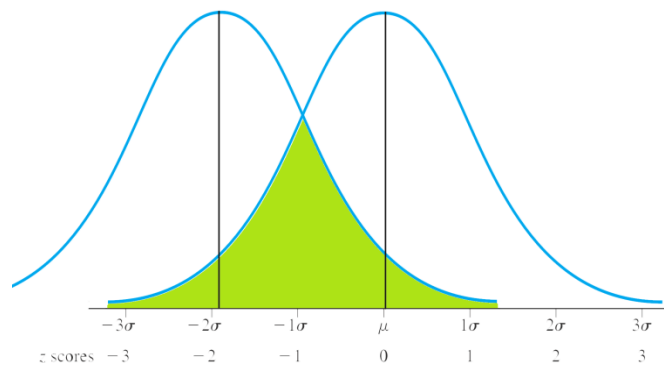


Figure 5: Comparing two example Normal distributions⁴.

The Kolmogorov-Smirnov test evaluates numerically the difference between two samples (e.g., F_1 and F_2) (Sheskin, 2004) or a sample and its reference distribution (e.g., F) (Papoulis and Pillai, 2002). The distance is the maximum vertical distance between two cumulative distribution functions given by:

$$d = \max_x |F_1(x) - F_2(x)|$$

or

$$d = \max_x |F_1(x) - F(x)|$$

respectively.

⁴ <http://www.researchgate.net/post/...>, accessed July 2013

4. Hypothesis Testing and Inferential Statistics

An experiment is a test under controlled conditions to investigate the validity of a hypothesis. A hypothesis gives a reasonable or educated guess to explain a phenomenon under investigation (Leedy and Ormrod, 2005, Salkind, 2008). According to the Collins Dictionary the term is derived from the Greek word *hupotithenai*, meaning to propose, suppose or put under. Proving a hypothesis is unfeasible because doing so would involve testing the entire population for validity (Salkind, 2008). Instead, hypothesis or significance testing involves experimenting on a sample and determining whether the results provide sufficient evidence to support the hypothesis. A well designed experiment seeks to mitigate sampling bias by ensuring samples approximate the characteristics of the population.

A variable in an experiment is any factor that can be introduced, changed, measured or controlled. Independent variables are factors that change in an experiment and are associated to the cause(s). Dependent variables relate to the effect(s) and are measured in an experiment. Controlled variables must be held constant throughout all treatments in an experiment. Allowing controlled variables to change causes them to become independent variables (Slavin, 2007). Confounding variables are additional, unaccounted for variables in an experiment. These confounding factors threaten the validity of conclusions made from an experiment because the treatment isn't the sole factor accounting for observed effects (Leedy and Ormrod, 2005).

A good hypothesis captures a problem statement or research question in a form that is more amenable to testing. Hypothesis testing involves studying the extent to which the independent variable (the cause or treatment) influences the dependent variable (the effect) (Leedy and Ormrod, 2005). The research or alternative hypothesis is a definitive statement of a relationship between variables. Here, the term variable refers to any quality or characteristic being investigated that has two or more possible values. The null hypothesis states the converse of the research hypothesis, that is no relationship exists between the variables (Salkind, 2008).

Descriptive statistics as described in Section 3 are techniques for describing a sample and include measures of central tendency (mean, median and mode), frequency distributions and graphs. While comparisons using this approach on data from two groups reveal differences between those specific groups, the results cannot be generalised to a larger population. Simply comparing means to determine treatment effects neglects to account for random variations due to chance. To account for these variations and for conclusions to be applied more generally require techniques from inferential statistics such as the Z-test, *t*-test, paired *t*-test, ANOVA and MANOVA described in this section. Non-parametric approaches include the Mann-Whitney or Wilcoxon signed-rank test.

Hypothesis testing determines whether the experimental evidence provides statistical support for the null hypothesis or its alternative by computing a *p*-value that gives the probability that the null hypothesis is wrong. Typically, the *p*-value is compared against a significance level of $\alpha = 0.05$ that requires the availability of moderate evidence for rejecting the null hypothesis in favour of the alternative (see Table 5).

Table 5: Interpreting p -values when conducting hypothesis testing.

p -value	Meaning
$p \geq 0.1$	No evidence against null hypothesis
$0.05 < p \leq 0.1$	Weak evidence against null hypothesis in favour of the alternative
$0.01 < p \leq 0.05$	Moderate evidence against null hypothesis in favour of the alternative
$0.001 < p \leq 0.01$	Strong evidence against null hypothesis in favour of the alternative
$p \leq 0.001$	Very strong evidence against null hypothesis in favour of the alternative

The value of α establishes the level of reasonable doubt, that is wrongly rejecting the null hypothesis when it is true. The confidence interval is evaluated by $1 - \alpha$. Setting $\alpha = 0$ implies never rejecting the null hypothesis.

4.1 Central Limit Theorem

Evidence to support or refute the null hypothesis is gathered by conducting multiple trials to assess the outcomes. We can represent the set of n outcomes as random variables X_1, \dots, X_n . That is, each trial is a random process that generates results according to some distribution, i.e. normal, bimodal, uniform, exponential... The term, *sample* is frequently overused in statistics and can either mean the outcome of a single trial or a set of n outcomes. In hypothesis testing, the term typically refers to the latter, with X_1, \dots, X_n being summarised by its mean value, \bar{X} .

One of the fundamental theorems in probability is the Central Limit Theorem (CLT) relating the distribution generating each trial outcome to the sampling distribution, \bar{X} . It states that if \bar{X} is the mean of n mutually independent random variables taken from any population with mean μ and standard deviation σ , then the probability distribution of:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (1)$$

tends to the standard normal probability distribution⁵ $N(0,1)$ as $n \rightarrow \infty$ (Smith, 1993). Similarly, the distribution of \bar{X} tends to $N(\mu, \sigma/\sqrt{n})$ as $n \rightarrow \infty$. The effect can be illustrated using convolution because if X and Y are independent random variables with density functions $f_X(x)$ and $f_Y(y)$ defined for all x and y , then the sum $Z = X + Y$ is a random variable with density (Grinstead and Snell, 1997, Papoulis and Pillai, 2002):

$$f_Z(z) = f_X(x) \otimes f_Y(y) \quad (2)$$

where convolution is defined as (Liu and Liu, 1975),

$$f(t) \otimes g(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau).d\tau. \quad (3)$$

Since

⁵ S. Khan (2011) *Central Limit Theorem*, http://www.khanacademy.org/math/probability/statistics-inferential/sampling_distribution, accessed April 2014

$$\begin{aligned}\bar{X} &= \frac{X_1 + X_2 + \dots + X_n}{n} \\ &= \frac{X_1}{n} + \frac{X_2}{n} + \dots + \frac{X_n}{n}\end{aligned}\quad (4)$$

then

$$f_{\bar{X}}(x) = f_X\left(\frac{x}{n}\right) \otimes \dots \otimes f_X\left(\frac{x}{n}\right). \quad (5)$$

For example, consider a stochastic process with a uniform distribution. The sampling distribution for $n = 2$ is a triangular distribution that can be generated by scaling the uniform distribution by a factor of 0.5 followed by convolving the resulting scaled distribution with itself. For any original distribution, as n increases, the sampling distribution increasingly resembles a normal distribution and is the basis for the CLT.

4.2 Z-Statistic

In some situations, the population's statistical properties are known so the experiment only involves testing the experimental group and assessing the probability (the p -value) that the results come from the population with known statistical properties. For example, body mass, height and intelligence quotient (IQ) scores have known population means and variances, enabling hypothesis testing using a single sample statistic. A significance test using the Z-statistic is useful for large values of n because the following statistic takes on a normal distribution:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}. \quad (6)$$

Here, \bar{X} gives the sample mean, $\mu_{\bar{X}}$ gives the mean of the sampling distribution and $\sigma_{\bar{X}}$ the standard deviation of the sampling distribution. Although the mean and standard deviation of the sampling distribution is often unknown, if we assume the null hypothesis that the sample was taken from the population, then $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ where μ and σ gives the mean and standard deviation of the population respectively. This allows the Z-statistic to take on the following form:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}. \quad (7)$$

According to the CLT, when the sample size is large ($n \rightarrow \infty$), samples from \bar{X} will appear normally distributed for arbitrary population distributions. However, when the population distribution appears normal, even when σ is unknown it is possible when $n \geq 20$ or 30 (Carlberg, 2011) to still use the Z-statistic by assigning $\sigma = s_{\bar{X}}$, the standard deviation of the sample means, for:

$$Z \approx \frac{\bar{X} - \mu}{\frac{s_{\bar{X}}}{\sqrt{n}}}; \quad \text{for large } n. \quad (8)$$

Most natural populations have a particular mathematical form that is termed normally distributed (Salkind, 2008).

The following example by Conley and Pollard⁶ is used to illustrate hypothesis testing using Z-statistics:

Suppose we are interested in the following research hypothesis:

H₁: The IQs of Yale students are higher than the general population.

After taking a random sample of 35 students and measuring their IQs, we find that their mean score is 107. Population IQ scores has a mean of 100 and a standard deviation of 15.

The first step involves assuming that the sample was taken from the general population; that is evaluating the probability of sampling $n = 35$ students whose mean IQ score is 107, comes from the general population. Leveraging off the CLT enables the standard deviation of the sample means for such a sample size to be computed:

$$\begin{aligned} s_{\bar{X}} &= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \\ &= \frac{15}{\sqrt{35}} = 2.535 \end{aligned}$$

This can then be used to work out a Z-score to evaluate how many standard deviations the result is from the mean:

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{\frac{s_{\bar{X}}}{\sqrt{n}}} \\ &= \frac{107 - 100}{2.535} = 2.761 \end{aligned}$$

That is, when sampling 35 students whose average IQ score equals 107 is 2.761 standard deviations away from the expected mean of 100 (from the CLT). Considering that we are dealing with a one-tailed test, our focus is in the area to left of the Z-score which in our case is 0.9971 (99.71%), in other words the sample score is higher than 99.71% of all other samples comprising of 35 students. The corresponding p -value is $1 - 0.9971 = 0.0029$ which is 29 chances out of 10,000 chances that this would occur if we assume that there is nothing special about the group.

In statistics, it is customary to use a significance level of $\alpha = 0.05$ corresponding to moderate evidence supporting the null hypothesis. This level of α implies a willingness to

⁶ D. Conley and D. Pollard (1998) *Hypothesis Testing, Confidence Intervals, and Power*, <http://www.yale.edu/soc119a/lecture7.htm>, accessed April 2013

be wrong once out of 20 times. For the example, for $\alpha = 0.05$, we are confident in rejecting the null hypothesis because the p -value of $0.0029 < \alpha$. In fact, strong evidence is available to support the alternative hypothesis.

4.3 Standard t -Test

Without understanding the population statistics (its mean and standard deviation), the experiment needs to be conducted on a control group without the treatment applied. The standard t -test is used to compare the means from exactly two groups and is the appropriate statistical test for such a case. It detects whether a statistical difference exists between the two group's results through a p -value expressing the probability that the null hypothesis is wrong. Assumptions for using the standard t -test include:

1. the two populations compared should both be Normally⁷ distributed
2. the two populations should have the same variance.

Whereas the Z -statistic takes on a normal distribution for large sample sizes, it is not the case for smaller n . That is, for smaller n , σ is not closely approximated by $s_{\bar{X}}$ and the resulting statistic takes on a t -distribution:

$$t = \frac{\bar{X} - \mu}{\frac{s_{\bar{X}}}{\sqrt{n}}}; \quad \text{for small } n. \quad (9)$$

The shape of the t -distribution depends on n : flatter for smaller values of n and taking the shape of the Normal distribution as the sample size increases (Carlberg, 2011). Its pdf is defined as follows (Papoulis and Pillai, 2002):

$$t(n) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n} \Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}} \quad \text{for } -\infty < x < \infty. \quad (10)$$

Here, the mean $\mu = 0$ and standard deviation is $\sigma = \sqrt{\frac{n}{n-2}}$ for $n > 2$.

After computing the t -statistic, the value is compared against a threshold from a t -test table giving the t -value required to reject the null hypothesis.

Population statistics are not always available, so a separate control group is used to estimate the population statistics. The t -statistic divides the difference between group means by the variation within and between the two groups (Salkind, 2008):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \left[\frac{n_1 + n_2}{n_1 n_2} \right]}}. \quad (11)$$

Here,

⁷ Tested using the Kolmogorov-Smirnov or Shapiro-Wilk test.

\bar{X}_1 :	Mean for group 1
\bar{X}_2 :	Mean for group 2
n_1 :	Number of participants in group 1
n_2 :	Number of participants in group 2
s_1^2 :	Variance for group 1
s_2^2 :	Variance for group 2

When the obtained value from Equation 11 is greater than the critical value from a standard t -table, then the null hypothesis can be rejected. Looking up the critical value t_{crit} from a t -table requires knowing the degrees of freedom df , the chosen level of significance α and whether a one or two-tailed test is being used. The t -test for independent means defines degrees of freedom as follows (Carlberg, 2011):

$$\begin{aligned} df &= n_1 - 1 + n_2 - 1 \\ &= n_1 + n_2 - 2 \end{aligned} \quad (12)$$

This allows the computed t -score to be referenced as $t_{(df)}$.

The following example illustrates applying a t -test to data obtained from two groups (Salkind, 2008).

A programme has been designed to help Alzheimer's patients remember the order of daily tasks. Group 1 was taught using visuals while group 2 was taught using visuals and intense verbal rehearsal. The data below counts the number of words remembered by each group.

Group 1						Group 2					
7	5	5	3	4	7	5	3	4	4	2	
3	6	1	2	10	9	4	5	2	5	4	
3	10	2	8	5	5	5	4	6	7	6	
8	1	2	5	1	12	8	7	8	8	7	
8	4	15	5	3	4	9	5	7	8	6	

Here, $\bar{X}_1 = 5.43$, $\bar{X}_2 = 5.53$, $s_1 = 3.42$, $s_2 = 2.06$, $n_1 = 30$ and $n_2 = 30$. The null and research hypotheses are as follows:

$$\mathbf{H}_0: \mu_1 = \mu_2$$

$$\mathbf{H}_1: \bar{X}_1 \neq \bar{X}_2$$

\mathbf{H}_1 is in the form of a two-tailed non-directional research hypothesis. The level of α (measuring accepted risk / Type 1 error / level of significance) is chosen to be 0.05. Here, the appropriate test is a t -test for independent means rather than one for dependent means because the groups are independent of one another while degrees of freedom $df = 30 + 30 - 2 = 58$:

$$\begin{aligned}
t_{(58)} &= \frac{5.43 - 5.53}{\sqrt{\left[\frac{(30-1)3.42^2 + (30-1)2.06^2}{30+30-2} \right] \left[\frac{30+30}{30 \times 30} \right]}} \\
&= \frac{-0.1}{\sqrt{\left(\frac{339.20 + 123.06}{56} \right) \left(\frac{60}{900} \right)}} \\
&= -0.18
\end{aligned}$$

For this problem, there is a requirement to identify the critical value where $df = 58$ and $\alpha = 0.05$ for a two-tailed t -test. Unfortunately, the standard tables for t -values do not give values for $df = 58$, only $df = 55$ or 60 . Salkind (2008) suggests choosing a degrees of freedom closest to that desired, in this case being $df = 60$. For this example, the obtained value to reject the null hypothesis needs to be equal or exceed $t_{crit} = 2.001$. The obtained value of -0.18 is less than 2.001 so the null hypothesis cannot be rejected. That is, the null hypothesis is the most attractive explanation.

4.4 Paired t -Test

Also referred to as the repeated measures t -test, the paired t -test is used when data has been collected from a single group of participants who are tested twice: before and after a treatment. While similar to the standard t -test, the availability of pre and post intervention data for each subject allows the two sets of results to be aggregated by subtracting one from the other. Here, the null hypothesis is demonstrated by showing that both sets of results come from the same population: $\mu_{post} = \mu_{pre}$ while assuming that $\sigma_{post} = \sigma_{pre}$. As with the standard t -test, assumptions for the paired t -test include:

1. the two populations compared should both be Normally⁸ distributed
2. the two populations should have the same variance.

The test statistic for the paired t -test is (Salkind, 2008):

$$t = \frac{\sum D}{\sqrt{\frac{n \sum D^2 - (\sum D)^2}{n-1}}} \quad (13)$$

where:

- $\sum D$: Sum of all differences between groups of scores
- $\sum D^2$: Sum of all differences squared between groups of scores
- n : Numbers of pairs of observations

⁸ Tested using the Kolmogorov-Smirnov or Shapiro-Wilk test.

As with the standard t -test, the paired t -test is also referred to by degrees of freedom $t_{(df)}$ although n in the following is referring to the pairs of observations (Carlberg, 2011):

$$df = n - 1 \quad (14)$$

allowing the computed t -score to be referenced as $t_{(df)}$. When the obtained value from Equation 13 is greater than the critical value obtained from a standard t -table, the null hypothesis can be rejected. The critical value is obtained from the cut-off for a t -statistic corresponding to a df value for some chosen level of significance α using a one or two-tailed test.

The following example from Salkind (2008) demonstrates hypothesis testing using a paired t -test:

Twenty-five participants were each tested before and after a treatment with the scores given in the table below.

	Participant																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Pre	3	5	4	6	5	5	4	5	3	6	7	8	7	6	7	8	8	9	9	8	7	7	6	7	8
Post	7	8	6	7	8	9	6	6	7	8	8	7	9	10	9	9	8	8	4	4	5	6	9	8	12

The null and research hypotheses are as follows:

$$H_0: \mu_{\text{post}} = \mu_{\text{pre}}$$

$$H_1: \bar{X}_{\text{post}} > \bar{X}_{\text{pre}}$$

as a one-tailed, directional research hypothesis. For this experiment, the level of significance or risk is chosen to be $\alpha = 0.05$ for a paired t -test. Testing the null hypothesis here is equivalent to testing whether $\mu_{\text{post}} - \mu_{\text{pre}} = 0$.

The difference and squared difference values between the two sets of scores are given below:

	Participant																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Pre	3	5	4	6	5	5	4	5	3	6	7	8	7	6	7	8	8	9	9	8	7	7	6	7	8
Post	7	8	6	7	8	9	6	6	7	8	8	7	9	10	9	9	8	8	4	4	5	6	9	8	12
Diff	4	3	2	1	3	4	2	1	4	2	1	-1	2	4	2	1	0	-1	-5	-4	-2	-1	3	1	4
Diff ²	16	9	4	1	9	16	4	1	16	4	1	1	4	16	4	1	0	1	25	16	4	1	9	1	16

Here, $\sum D = 30$, $\sum D^2 = 180$, $n = 25$. Applying the t -test given in Equation 13 gives the following result:

$$\begin{aligned}
 t_{(24)} &= \frac{30}{\sqrt{\frac{(25 \times 180) - 30^2}{25 - 1}}} \\
 &= \frac{30}{\sqrt{150}} \\
 &= 2.45
 \end{aligned}$$

The critical value for the t -test for a one-tailed test, $df = 25 - 1 = 24$ and $\alpha = 0.05$ is 1.711 using a standard t -table. Here, the null hypothesis can be rejected since the obtained value $t_{(24)} = 2.45$ is greater than the critical value 1.711. That is, the result is extreme enough according to our level of accepted risk that the difference between the pre and post-treatment results occurred not by chance but was due to the treatment.

4.5 ANOVA

Previously, we examined the two-sample t -test used to compare two population means μ_1 and μ_2 . Analysis of Variance (ANOVA) is a hypothesis testing technique testing the equality of two or more population means. Often known as one-way ANOVA for comparing the means of more than two groups or levels of an independent variable, the analysis involves comparing the variances of samples taken due to differences between individuals within groups and between groups (Salkind, 2008). The same assumptions for the t -test apply to ANOVA and include (Smith, 1993):

1. normally distributed data
2. independent samples taken from each of the treatments
3. the population standard deviation σ is common to each treatment.

ANOVA generalises the standard t -test to enable comparisons of $k > 2$ population means. As a result, the null hypothesis becomes (Smith, 1993),

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

against an alternative hypothesis that at least two of the population means differ. Central to a $k > 2$ comparison of means is the following total sum of squares (SS_{total}):

$$SS_{\text{total}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{..})^2 \quad (15)$$

Here, $\bar{X}_{..}$ denotes the grand mean of all samples, i references a particular observation given by X_{ij} in treatment j while n_j gives the total number of observations in treatment j . Adding and subtracting \bar{X}_j does not affect the final sum but allows SS_{total} to expand to⁹,

$$\begin{aligned} SS_{\text{total}} &= \sum_{j=1}^k \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X}_{..})]^2 \\ &= \sum_{j=1}^k \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_j)^2 + 2(X_{ij} - \bar{X}_j)(\bar{X}_j - \bar{X}_{..}) + (\bar{X}_j - \bar{X}_{..})^2] \\ &= \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 + 2 \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)(\bar{X}_j - \bar{X}_{..}) + \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X}_{..})^2 \end{aligned} \quad (16)$$

⁹ K. McIntyre (2005) *Notes on ANOVA*, http://ww2.mcdaniel.edu/Bus_Econ/mcintyre/ANOVA.PDF, accessed April 2013

The middle term here can be crossed off because deviations from the mean always sum to zero. Since the third term makes no references to X_{ij} , SS_{total} can be written as follows (Smith, 1993),

$$\begin{aligned} SS_{\text{total}} &= \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 + \sum_{j=1}^k n_j (\bar{X}_j - \bar{X}_{..})^2 \\ &= SS_{\text{residuals}} + SS_{\text{treatments}} \end{aligned} \quad (17)$$

If s_j gives the sample standard deviation of the j^{th} sample, then $SS_{\text{residuals}}$ can be expressed as,

$$\begin{aligned} SS_{\text{residuals}} &= \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \\ &= \sum_{j=1}^k (n_j - 1) s_j^2 \end{aligned} \quad (18)$$

while

$$SS_{\text{treatments}} = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X}_{..})^2. \quad (19)$$

$SS_{\text{treatments}}$ measures the differences between groups while $SS_{\text{residuals}}$ captures the differences between individuals within groups. The following provide two choices of two unbiased mean square estimators of σ^2 :

$$MS_{\text{treatments}} = \frac{SS_{\text{treatments}}}{df_{\text{treatments}}} \quad (20)$$

where,

$$df_{\text{treatments}} = k - 1 \quad (21)$$

and

$$MS_{\text{residuals}} = \frac{SS_{\text{residuals}}}{df_{\text{residuals}}} \quad (22)$$

where,

$$df_{\text{residuals}} = n - k. \quad (23)$$

where n gives the total number of measures,

$$n = \sum_{j=1}^k n_j \quad (24)$$

The total degrees of freedom,

$$df_{\text{total}} = n - 1 \quad (25)$$

An F -test is used to check whether the variance between treatments is the same as the variance between individuals within treatments are equal. Using ANOVA,

$$F = \frac{MS_{\text{treatments}}}{MS_{\text{residuals}}} \quad (26)$$

The F -value indicates how far the data doesn't support the null hypothesis as a deviation from $F = 1$. A large value of F implies that the effect of the treatment is relevant because the variability between samples is much larger than the variability within each sample. The critical value for the F -test can be found from an F -table and requires knowing the Type 1 error rate α , $df_{\text{treatments}}$ and $df_{\text{residuals}}$. The null hypothesis is rejected when the obtained F -value is larger than the critical value F_{crit} . The computed values are then summarised in an ANOVA table:

Table 6: Composition of a one-way ANOVA table

Source	Degrees of Freedom	Sums of Squares	Mean Squares	F	F_{crit}
Treatments (between)	$df_{\text{treatments}}$	$SS_{\text{treatments}}$	$MS_{\text{treatments}}$	$MS_{\text{treatments}} / MS_{\text{residuals}}$	F_{crit}
Residuals (within)	$df_{\text{residuals}}$	$SS_{\text{residuals}}$	$MS_{\text{residuals}}$		
Total	df_{total}	SS_{total}			

The following example from Smith (1993) illustrates applying the one-way ANOVA to support hypothesis testing:

Samples of 25 kg packs of PVC powder were selected from 5 batches coming off an assembly line. The discrepancy between the label weight and its weighed value was noted in units of $\times 10^{-2}$ kg. It is assumed that the 5 samples were independent, normally distributed with a common variance. The recorded data is as follows:

Table 7: Measurements from batch example.

		Batch ($j \in [1, k]$) ... (Treatments)				
		1	2	3	4	5
$i \in [1, n]$... (Measures)	1	0	6	7	-25	1
	2	-15	-28	-13	-45	10
	3	8	5	-13	17	-45
	4	36	5	-40	1	-15
	5	7	-19	15	-14	-3
	6	27	-10	-15	1	-28
	7	10	25	17	-5	30
	8	15	-14	-38	-50	11
	9	15	10	-5	10	-5
	10	24	0	-2	-35	-50
	11	29	15	-37	10	9
M_j :		14.18	-0.45	-11.27	-12.27	-7.73
GM:		-3.51	-3.51	-3.51	-3.51	-3.51
$M_j - GM$:		17.69	3.05	-7.76	-8.76	-4.22
SD $_j$:		14.55	15.73	20.42	23.30	24.73

The terms referred to as M_j are the means calculated over a single sample,

$$\bar{X}_j = \frac{\sum_{i=1}^n X_{ij}}{n}$$

such that for example,

$$\bar{X}_1 = \frac{0 - 15 + 8 + 36 + 7 + 27 + 10 + 15 + 15 + 24 + 29}{11} = 14.18.$$

The grand mean (GM) is given by,

$$\begin{aligned}\bar{X}_{..} &= \frac{\sum_{j=1}^k \sum_{i=1}^n X_{ij}}{k \times n} \\ &= \frac{0 - 15 + 8 + \dots - 5 - 50 + 9}{5 \times 11} \\ &= \frac{-193}{55} = -3.51\end{aligned}$$

The standard deviation referred to in the table as SD_j are calculated over a single sample,

$$s_j = \sqrt{\frac{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}{n - 1}}$$

such that for example,

$$\begin{aligned}s_1 &= \sqrt{\frac{(0 - 14.18)^2 + (-15 - 14.18)^2 + \dots + (24 - 14.18)^2 + (29 - 14.18)^2}{11 - 1}} \\ &= \sqrt{\frac{2117.636}{10}} = 14.55\end{aligned}$$

The ANOVA table for this example is,

Table 8: ANOVA table for batch example.

Source	Degrees of Freedom	Sums of Squares	Mean Squares	<i>F</i>	<i>F</i> _{crit}
Treatments (between)	4	5248.84	1312.21	3.23	2.56
Residuals (within)	50	20306.91	406.14		
Total	54	25555.75			

$SS_{\text{treatments}}$ and $SS_{\text{residuals}}$ are computed as follows,

$$\begin{aligned}SS_{\text{treatments}} &= \sum_{j=1}^k n_j (\bar{X}_j - \bar{X}_{..})^2 \\ &= 11 \times 17.69^2 + 11 \times 3.05^2 + 11 \times (-7.76)^2 + 11 \times (-8.76)^2 + 11 \times (-4.22)^2 \\ &= 5,248.84\end{aligned}$$

and

$$\begin{aligned}
SS_{\text{residuals}} &= \sum_{j=1}^k (n_j - 1) s_j^2 \\
&= 10 \times 14.55^2 + 10 \times 15.73^2 + 10 \times 20.42^2 + 10 \times 23.30^2 + 10 \times 24.73^2 \\
&= 20,306.91
\end{aligned}$$

SS_{total} is therefore,

$$\begin{aligned}
SS_{\text{total}} &= SS_{\text{residuals}} + SS_{\text{treatments}} \\
&= 5,248.84 + 20,306.91 \\
&= 25,555.75
\end{aligned}$$

The degrees of freedom are calculated as follows,

$$\begin{array}{lll}
df_{\text{treatments}} = k - 1 & df_{\text{residuals}} = n - k & df_{\text{total}} = n - 1 \\
= 5 - 1 & = 55 - 5 & = 55 - 1 \\
= 4 & = 50 & = 54
\end{array}$$

with mean squares,

$$\begin{array}{ll}
MS_{\text{treatments}} = \frac{SS_{\text{treatments}}}{df_{\text{treatments}}} & MS_{\text{residuals}} = \frac{SS_{\text{residuals}}}{df_{\text{residuals}}} \\
= \frac{5,248.84}{4} & = \frac{20,306.91}{50} \\
= 1,312.21 & = 406.14
\end{array}$$

The obtained value for the F -value is,

$$\begin{aligned}
F &= \frac{MS_{\text{treatments}}}{MS_{\text{residuals}}} \\
&= \frac{1,312.21}{406.14} \\
&= 3.23
\end{aligned}$$

F_{crit} for Type 1 error of $\alpha = 0.05$, $df_{\text{treatments}} = 4$ and $df_{\text{residuals}} = 50$ is 2.56.

Since the obtained value of 3.23 is larger than the critical value of 2.56, then the evidence supports rejecting the null hypothesis – at least two of the batches are significantly different. The box plot in Figure 6 shows the quartiles of each sample with the whiskers representing the smallest and largest values observed (Coakes and Ong, 2011). The diagram shows that batch 1 is the likely cause of the rejection of H_0 .

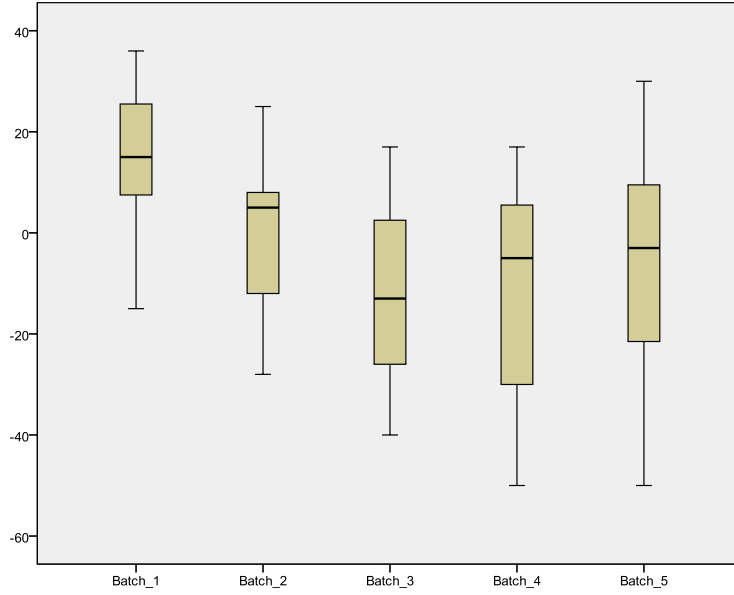


Figure 6: Box and whisker plot of the batch data produced with SPSS.

A closer look at Figure 6 shows that batch 2 is highly skewed and requires a further assessment of normality. With $n_2 = 11$ (less than 100) a test involving the Kolmogorov-Smirnov and Shapiro-Wilk statistics is used to confirm this requirement (Coakes and Ong, 2011). Employing SPSS produces significance values of 0.2 and 0.896 respectively, both exceeding 0.05 and thus confirming the assumption of normality.

4.6 Two-Way ANOVA

Section 4.5 detailed the one-way ANOVA for observations of a single dependent variable influenced by a single independent variable (the treatment). Cases involving multiple independent variables affecting a single dependent variable involve factorial analysis of variance (factorial ANOVA). The two-way ANOVA is the simplest kind of factorial ANOVA and is used when there are two independent variables (Salkind, 2008). The two independent variables typically comprise of a factor and treatment producing an outcome measured by the dependent variable. A two-way ANOVA firstly applies a one-way ANOVA on the independent variables, followed by an analysis of whether both factors together affect the outcome. The total sum of squares for a two-way ANOVA with independent variables A and B is (Moore and McCabe, 2003, Sparks, 2011),

$$\begin{aligned}
 SS_{\text{total}} &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{X} \dots) \\
 &= SS_A + SS_B + SS_{AB} + SS_{\text{residuals}}
 \end{aligned} \tag{27}$$

Here, we have avoided labelling A and B as treatments because often just one variable is a treatment while the other represents a factor. Consider factors A and B as having a and b levels respectively, and n_{ij} observations for $i \in [1, a]$ and $j \in [1, b]$. The total number of observations is therefore,

$$n = \sum_{i=1}^a \sum_{j=1}^b n_{ij} . \quad (28)$$

SS_A , SS_B and SS_{AB} and $SS_{residuals}$ in Table 9 are computed as by ignoring the unfocussed variable and performing a one-way ANOVA on the data on A and B respectively as follows

$$SS_A = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{X}_{i..} - \bar{X}_{...})^2 , \quad (29)$$

$$SS_B = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{X}_{.j.} - \bar{X}_{...})^2 , \quad (30)$$

$$SS_{AB} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2 \quad (31)$$

and

$$\begin{aligned} SS_{residuals} &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{ij.})^2 \\ &= \sum_{i=1}^a \sum_{j=1}^b (n_{ij} - 1) s_{ij}^2 \end{aligned} \quad (32)$$

where s_{ij} denotes the standard deviation within group i, j .

Table 9: Composition of a two-way ANOVA table

	Sums of				
Source	Degrees of Freedom	Squares	Mean Squares	F	F _{crit}
Factor A (between)	$df_A = a - 1$	SS_A	$MS_A = SS_A / df_A$	$MS_A / MS_{residuals}$	$F_{crit,A}$
Factor B (between)	$df_B = b - 1$	SS_B	$MS_B = SS_B / df_B$	$MS_B / MS_{residuals}$	$F_{crit,B}$
Factor AB (between)	$df_{AB} = (a - 1)(b - 1)$	SS_{AB}	$MS_{AB} = SS_{AB} / df_{AB}$	$MS_{AB} / MS_{residuals}$	$F_{crit,AB}$
Residuals (within)	$df_{residuals} = n - ab$	$SS_{residuals}$	$MS_{residuals} = SS_{residuals} / df_{residuals}$		
Total	$df_{total} = n - 1$	SS_{total}			

To simplify the calculations, the statistics package SPSS has been used to work through the following two-way ANOVA example from Salkind (2008).

A study involves studying the impact of gender and exercise program (two independent variables) on weight loss (the dependent variable). The experimental design is as follows:

		Exercise Program	
		High Impact	Low Impact
Gender	Male		
	Female		

The data collected from the experiment is as follows:

Treatment:	High Impact		Low Impact	
Gender:	Male	Female	Male	Female
	76	65	88	65
	78	90	76	67
	76	65	76	67
	76	90	76	87
	76	65	56	78
	74	90	76	56
	74	90	76	54
	76	79	98	56
	76	70	88	54
	55	90	78	56

A two-way ANOVA produces three null hypotheses. The first assesses the impact of the exercise program on the outcome of weight loss,

$$\mathbf{H}_0: \mu_{\text{high}} = \mu_{\text{low}}$$

$$\mathbf{H}_1: \bar{X}_{\text{high}} \neq \bar{X}_{\text{low}}.$$

The second assesses whether gender has an impact on weight loss,

$$\mathbf{H}_0: \mu_{\text{male}} = \mu_{\text{female}}$$

$$\mathbf{H}_1: \bar{X}_{\text{male}} \neq \bar{X}_{\text{female}}.$$

The third hypothesis assesses whether there's an interaction effect of exercise program and gender on treatment on weight loss,

$$\mathbf{H}_0: \mu_{\text{high,male}} = \mu_{\text{high,female}} = \mu_{\text{low,male}} = \mu_{\text{low,female}}$$

$$\mathbf{H}_1: \bar{X}_{\text{high,male}} \neq \bar{X}_{\text{high,female}} \neq \bar{X}_{\text{low,male}} \neq \bar{X}_{\text{low,female}}.$$

The application of SPSS on the data produces the output

Table 10: Applying a two-way ANOVA in SPSS on the data.

Tests of Between-Subjects Effects

Dependent Variable: Loss

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	218892.025	1	218892.025	1057.322	.020
	Error	207.025	1	207.025 ^a		
Treatment	Hypothesis	265.225	1	265.225	.252	.704
	Error	1050.625	1	1050.625 ^b		
Gender	Hypothesis	207.025	1	207.025	.197	.734
	Error	1050.625	1	1050.625 ^b		
Treatment * Gender	Hypothesis	1050.625	1	1050.625	9.683	.004
	Error	3906.100	36	108.503 ^c		

a. MS(Gender)

b. MS(Treatment * Gender)

c. MS(Error)

The first hypothesis is tested as though the data excludes gender information and a one-way ANOVA performed on the treatment of exercise program. From Table 10, SPSS returns a p -value of 0.704 (under the Sig. column) which is larger than $\alpha = 0.05$, meaning we fail to reject the null hypothesis. That is, the choice of exercise program (either high or low impact) by itself has no effect on weight loss.

The second hypothesis is tested on the data while ignoring treatment information by performing a one-way ANOVA on gender. The analysis results in a p -value of 0.734 for gender being larger than $\alpha = 0.05$, meaning we fail to reject the null hypothesis. We conclude that gender (being either male or female) in isolation has no effect on weight loss.

The third hypothesis is tested by considering whether both independent variables of gender and exercise program have an effect on weight loss. The p -value of 0.004 under the final column of the row indicated by Treatment * Gender is less than $\alpha = 0.05$, thus indicating that both factors affect the outcome. In other words, an interaction exists between the explanatory variables (Seltman, 2013) and we conclude that the effect of changes in the exercise program depends on the gender (the other factor).

4.7 MANOVA

Multivariate analysis of variance (MANOVA) extends ANOVA by enabling the analysis of multiple dependent variables (DVs). Remember that ANOVA tests for the truth of the null hypothesis that the means of k treatments are equal:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k . \quad (33)$$

MANOVA tests for the truth of the null hypothesis that k treatments with a p -dimensional vector of population means are equal (Bray and Maxwell, 1985):

$$\begin{aligned} & \mu_{11} = \mu_{12} = \dots = \mu_{1k} \\ H_0 : & \begin{array}{c} \mu_{21} = \mu_{22} = \dots = \mu_{2k} \\ \vdots \\ \mu_{p1} = \mu_{p2} = \dots = \mu_{pk} \end{array}, \end{aligned} \quad (34)$$

allowing relationships between DVs to be assessed. Whereas ANOVA only uses the F -test to form a test statistic, MANOVA uses several including:

- Wilks' lambda
- Pillai-Bartlett trace
- Roy's greatest characteristic root
- Hotelling-Lawley trace.

These statistics are derived from the full and reduced models of errors associated with the data X_{ij} for trial i in treatment j . For the full model, the individual error estimate is given by,

$$\hat{\varepsilon}_{ij}(F) = X_{ij} - \bar{X}_j \quad (35)$$

and is applied to each DV. The sum of squared errors given by,

$$\begin{aligned} SSE(F) &= \sum_j \sum_i \hat{\varepsilon}_{ij}^2(F) \\ &= \sum_j \sum_i (X_{ij} - \bar{X}_j)^2 \end{aligned} \quad (36)$$

is also referred to as the sum of squares within treatments $SS_{\text{residuals}}$. Similarly for the reduced model, the individual error estimate given by,

$$\hat{\varepsilon}_{ij}(R) = X_{ij} - \bar{X}_\bullet \quad (37)$$

is also applied to each DV. The sum of squared errors given by,

$$\begin{aligned} SSE(R) &= \sum_j \sum_i \hat{\varepsilon}_{ij}^2(R) \\ &= \sum_j \sum_i (X_{ij} - \bar{X}_\bullet)^2 \end{aligned} \quad (38)$$



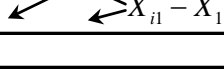
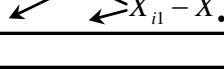


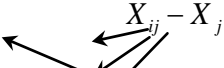
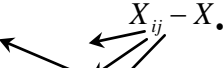
is also called the total sum of squares, SS_{total} . The sum of squares between treatments,

$$\begin{aligned} SS_{\text{treatments}} &= SS_{\text{total}} - SS_{\text{residuals}} \\ &= n_j (X_j - \bar{X}_\bullet)^2 \end{aligned} \quad (39)$$

where n_j represents the number of trials in treatment j .

For simplicity, the remainder of this section will focus on the two DV based MANOVA. Table 11 illustrates computing the errors for full and reduced models for such a problem. For simplicity, errors associated with DV₁ are labelled under e_1 and those related to DV₂ labelled under e_2 for both models.

Table 11: Errors for Full and Reduced Models for two-way MANOVA data

Treatment	Trial	Full Model					Reduced Model				
		e_1	e_2	$(e_1)^2$	$(e_2)^2$	$e_1 e_2$	e_1	e_2	$(e_1)^2$	$(e_2)^2$	$e_1 e_2$
1	1										
	...										
	i										
	Sum:										
...	1										
	...										
	i										
	Sum:										
j	1										
	...										
	i										
	Sum:										
Grand Sum:											

The F -test is the basis of the univariate test of significance, comparing $\sum_j \sum_i e_1^2(F)$ against $\sum_j \sum_i e_1^2(R)$ for DV₁ and separately for all other DVs. MANOVA also considers the relationship between DVs by also comparing $\sum_j \sum_i e_1 e_2(F)$ against $\sum_j \sum_i e_1 e_2(R)$. These sums of cross-products is closely related to the correlation between the two variables where,

$$r_{Y_1 Y_2}(\text{within treatments}) = \frac{\sum_j \sum_i e_1 e_2(F)}{\sqrt{\sum_j \sum_i e_1^2(F) \sum_j \sum_i e_2^2(F)}} \quad (40)$$

and

$$r_{Y_1 Y_2}(\text{total sample}) = \frac{\sum_j \sum_i e_1 e_2(R)}{\sqrt{\sum_j \sum_i e_1^2(R) \sum_j \sum_i e_2^2(R)}}. \quad (41)$$

The inclusion of cross product terms for MANOVA results in the following error matrix for the full model,

$$\mathbf{E} = \begin{bmatrix} \sum_j \sum_i e_1^2(\mathbf{F}) & \sum_j \sum_i e_1 e_2(\mathbf{F}) \\ \sum_j \sum_i e_1 e_2(\mathbf{F}) & \sum_j \sum_i e_2^2(\mathbf{F}) \end{bmatrix} \quad (42)$$

and the following error matrix for the reduced model,

$$\mathbf{T} = \begin{bmatrix} \sum_j \sum_i e_1^2(\mathbf{R}) & \sum_j \sum_i e_1 e_2(\mathbf{R}) \\ \sum_j \sum_i e_1 e_2(\mathbf{R}) & \sum_j \sum_i e_2^2(\mathbf{R}) \end{bmatrix}. \quad (43)$$

The hypothesis sum of squares and cross-product matrix is given by, $\mathbf{H} = \mathbf{T} - \mathbf{E}$. The four multivariate test statistics for MANOVA referred to above are all based on eigenvalues λ_i of $\mathbf{H}\mathbf{E}^{-1}$ where $i = 1, \dots, s$. The test statistic for Wilks' lambda is,

$$U = \prod_{i=1}^s \frac{1}{1 + \lambda_i}. \quad (44)$$

The formula for the Pillai-Bartlett trace is given by,

$$V = \sum_{i=1}^s \frac{\lambda_i}{1 + \lambda_i} \quad (45)$$

while the test statistic for Roy's greatest characteristic root is,

$$\text{GCR} = \frac{\lambda_1}{1 + \lambda_1}. \quad (46)$$

The formula for the Hotelling-Lawley trace is,

$$T = \sum_{i=1}^s \lambda_i. \quad (47)$$

Determining whether the null hypothesis should be rejected requires comparing the observed value of the test statistic to the sampling distribution of the statistic under the null hypothesis. This requires a transformation of each test statistic to a variable approximating the F -distribution with the intention of deriving a p -value. As SPSS produces the desired outputs automatically, we won't be detailing the calculations. However, Bray and Maxwell (1985) summarise the transformations from the U and V -statistics to those approximating an F -variable for those interested in the calculations. The resulting p -values are compared to the chosen significance level α for accepting or rejecting the null hypothesis.

4.8 Mann-Whitney U -Test / Wilcoxon Signed-Rank Test

The Mann-Whitney U -test is considered the non-parametric equivalent of the independent samples t -Test and can be performed on ordinal (ranked) data. The Mann-Whitney U -test is well suited to Likert item data as it cannot be presumed that the underlying population

fits a normal distribution and the Mann-Whitney U -test is included in most modern statistical packages. The analysis of individual Likert items will be in terms of determining if there is a statistical difference in responses between the two treatments.

Although the Mann-Whitney test does not require normally distributed data it does require that the data from each population must be an independent random sample, and the population distributions must have equal variances and the same shape. Equal variance can be tested through the application of a non-parametric version of the Levine's test for homogeneity of variance. This test can also be used to support required assumption that the two samples come from the same distribution shape. In addition, a visual inspection of the probability plot may help in determining if the distributions look similar.

Testing the Likert item results consist of creating a null and alternative hypothesis. The null hypothesis, H_0 is: The samples come from the same distribution, or there is no difference in ranks between the treatments. The alternative hypothesis, H_1 is: The samples come from different distribution, or there is a significant difference, typically $\alpha = 0.05$.

The Mann Whitney U -statistic is defined as¹⁰:

$$U = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum_{i=n_1+1}^{n_2} R_i \quad (48)$$

where R_i are the rank values.

The U -test can be calculated by hand for small data sets. The calculation procedure is conducted as described below:

1. Rank the results, where there is a tie use the average value.
2. Add up the ranks for the observations which came from sample 1 (R_1) and use the expression below to calculate U_1 :

$$U_1 = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \quad (49)$$

For sample size N_1 and sum of ranks R_1 in sample 1.

3. Use the expression below to calculate U_2 ,

$$U_2 = N_1 N_2 - U_1 \quad (50)$$

Use the smaller value of U_1 and U_2 when consulting significance tables. Note: The Mann-Whitney U -statistic follows a Z-distribution when the sample size is greater than 20. For values less than 20 refer to the Mann-Whitney Critical Value table.

¹⁰ Stats Direct, Mann Whitney U-Test,

www.statsdirect.com/webhelp/#nonparametric_methods/mann_whitney.htm, accessed April 2014

5. Summary

A hypothesis has to be capable of being tested by an experiment. Well-designed experiments provide a systematic approach to observe relations among variables while ruling out alternative explanations. This report has examined a number of experimental designs including the simple experiment, matched-pairs, repeated-measures and the single group design. While uncontrolled experiments such as those conducted in the field do not mitigate all confounding factors, they provide external-validity to support the results from controlled experiments.

A difficulty with human-in-the-loop experiments is the requirement to gauge behaviours, characteristics, attitudes and opinions. An approach to generate quantitative data from such responses is to use Likert scales. Once collected, the data can be analysed using descriptive statistics or compared using a number of approaches to visualise results. These approaches reveal differences between the specific groups but to generalise the results to a larger population requires inferential statistics.

6. References

- Bray, J. H. and Maxwell, S. E. (1985) *Multivariate Analysis of Variance*. Beverly Hills, USA, Sage Publications
- Carlberg, C. (2011) *Statistical Analysis: Microsoft Excel 2010*. Indianapolis, Que Publishing
- Coakes, S. J. and Ong, C. (2011) *SPSS: Analysis without Anguish (version 18.0 for Windows)*. Milton, Australia, Wiley
- Few, S. (2005) Keep Radar Graphs Below the Radar – Far Below. *DM Review* **15** (5) 48
- Grinstead, C. M. and Snell, J. L. (1997) *Introduction to Probability*. 2nd revised ed. Rhode Island, USA, American Mathematical Society
- Gueorguieva, R. and Krystal, J. H. (2004) Move over ANOVA. *Archives of General Psychiatry* **61** 310–317
- Kass, R. A. (2006) *The Logic of Warfighting Experiments*. Washington, USA, CCRP Publications
- Kass, R. A., et al. (2006) *Guide for Understanding and Implementing Defense Experimentation (GUIDEx)*. Ottawa, Canada, TTCP
- Leedy, P. D. and Ormrod, J. E. (2005) *Practical Research - Planning and Design*. 8th ed. New York, Pearson Education
- Liu, C. L. and Liu, J. W. S. (1975) *Linear Systems Analysis*. Tokyo, Japan, McGraw-Hill Kogakusha
- Lo, E. H. S., et al. (2013) *Analysis of Team Interactions in the AOC during Exercise Pitch Black 2008*. DSTO-RR-0384, [Research Report] Canberra, Australia, DSTO
- Lo, E. H. S., Au, T. A. and Hoek, P. J. (2014) Improving the Design of a Human-in-the-Loop Joint Fires Experiment. In: *Proc. of Int'l Conf. on System of Systems Engineering*, Adelaide, Australia: June, IEEE
- Madsen, K., Nielsen, H. B. and Tingleff, O. (2004) *Methods for Non-Linear Least Square Problems*. Lyngby, Denmark, Technical University of Denmark
- Mergerdichian, B., et al. (2012) *CK-12 Middle School Math Grade 6*. Vol. 1, CK-12 Foundation
- Mitchell, M. L. and Jolley, J. M. (2010) *Research Design Explained*. 7th ed. Belmont, USA, Wadsworth
- Moore, D. S. and McCabe, G. P. (2003) *Introduction to the Practice of Statistics*. New York, W. H. Freeman & Co.
- Neuman, W. L. (2006) *Social Research Methods*. 6th ed. Boston, USA, Pearson Education
- Nock, M. K., Michel, B. D. and Photos, V. I. (2008) Single-Case Research Designs. In: McKay, D. (ed.) *Handbook of Research Methods in Abnormal and Clinical Psychology*. Beverly Hills, USA, SAGE Publications 337–350
- Papoulis, A. and Pillai, S. U. (2002) *Probability, Random Variables, and Stochastic Processes*. 4th ed. New York, McGraw-Hill

- Salkind, N. J. (2008) *Statistics for People Who (Think They) Hate Statistics*. 3rd ed. Los Angeles, SAGE Publications
- Seltman, H. (2013) *Experimental Design and Analysis*. Pittsburgh, USA, Carnegie Mellon University
- Sheskin, D. J. (2004) *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd ed. Boca Raton, USA, CRC Press
- Slavin, R. E. (2007) *Educational Research in an Age of Accountability*. New York, USA, Pearson Education
- Smith, P. J. (1993) *Into Statistics*. Melbourne, Thomas Nelson
- Sparks, D. (2011) Basics of Two-Way ANOVA. In: *Introduction to Statistics 2*. Gainesville, University of Florida

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA					
				1. DLM/CAVEAT (OF DOCUMENT)	
2. TITLE A Short Guide to Experimental Design and Analysis for Engineers			3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION) Document (U) Title (U) Abstract (U)		
4. AUTHOR(S) Edward H. S. Lo, T. Andrew Au and Peter J. Hoek			5. CORPORATE AUTHOR DSTO Defence Science and Technology Organisation 506 Lorimer St Fishermans Bend Victoria 3207 Australia		
6a. DSTO NUMBER DSTO-TN-1291		6b. AR NUMBER AR-015-938		6c. TYPE OF REPORT Technical Note	
				7. DOCUMENT DATE April 2014	
8. FILE NUMBER -	9. TASK NUMBER 07-294	10. TASK SPONSOR DG Joint Force Integration		11. NO. OF PAGES 34	12. NO. OF REFERENCES 25
13. DSTO Publications Repository http://dspace.dsto.defence.gov.au/dspace/			14. RELEASE AUTHORITY Chief, Joint and Operations Analysis Division		
15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT <i>Approved for public release</i> OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111					
16. DELIBERATE ANNOUNCEMENT No Limitations					
17. CITATION IN OTHER DOCUMENTS Yes					
18. DSTO RESEARCH LIBRARY THESAURUS Human-in-the-loop; experimentation; statistical inference					
19. ABSTRACT An experiment is a test under controlled conditions to investigate the validity of a hypothesis. Experimentation is the basis for creating new knowledge – determining whether some factor causes an effect. Well-designed experiments provide a systematic approach to observe relations among variables while ruling out alternative explanations. This report has examined a number of experimental designs including the simple experiment, matched-pairs, repeated-measures and the single group design. The relevant statistical techniques are also discussed to help identify key quantitative methods for data processing and analysis. While not intending to replace any textbook, this guide summarises the resources and some worked examples that have helped the authors gain a basic understanding of design, measurement and statistical analysis to support military experiments.					